A comparison of central-tendency and interconnectivity approaches to clustering multivariate data with irregular structure

Mark Tozer[1] and David Keith[1]

[1]University of New South Wales

September 6, 2022

Abstract

Abstract Questions: Most clustering methods assume data are structured as discrete hyper-spheroidal clusters to be evaluated by measures of central-tendency. If vegetation data do not conform to this model, then vegetation data may be clustered incorrectly. What are the implications for cluster stability and evaluation if clusters are of irregular shape or density? Location: Southeast Australia Methods: We define misplacement as the placement of a sample in a cluster other than (distinct from) its nearest neighbour and hypothesise that optimising homogeneity incurs the cost of higher rates of misplacement. The Chameleon algorithm emphasises interconnectivity and thus is sensitive to the shape and distribution of clusters. We contrasted its solutions with those of traditional non-hierarchical and hierarchical (agglomerative and divisive) approaches. Results: Chameleon-derived solutions had lower rates of misplacement and only marginally higher heterogeneity than those of k-means in the range 15–60 clusters, but their metrics converged with larger numbers of clusters. Solutions derived by agglomerative clustering had the best metrics (and divisive clustering the worst) but both produced inferior high-level solutions clusters to those of Chameleon by merging distantly-related clusters. Conclusions: Our results suggest that Chameleon may have an advantage over traditional algorithms at when data exhibit discontinuities and variable structure, potentially producing more stable solutions (due to lower rates of misplacement), but scoring lower on traditional metrics of central-tendency. Chameleon's advantages are less obvious in the partitioning of data from continuous gradients, however its graph-based partitioning protocol facilitates hierarchical integration of solutions.

A comparison of central-tendency and interconnectivity approaches to clustering multivariate data with irregular structure

[1]Mark Tozer and David Keith

Centre for Ecosystem Science, School of Biological, Earth and Environmental Science, University of NSW, Sydney, NSW 2052, Australia

[1] Corresponding author (Email: m.tozer@student.unsw.edu.au)

Suggested running title:

Dealing with irregular data structure

Keywords

Chameleon, classification, misclassification, cluster-optimisation, clustering, CLUTO, vegetation databases, cluster metrics, homogeneity

Abstract

Questions: Traditional clustering methods generally assume data are structured as discrete hyper-spheroidal clusters to be evaluated by measures of central-tendency. If vegetation data do not conform to this model, then vegetation data may be clustered incorrectly. What are the implications for cluster stability and evaluation if clusters are of irregular shape or density?

Location: Southeast Australia

Methods: We define mis-classification as the placement of a sample in a cluster other than its nearest neighbour and hypothesise that: i) optimising homogeneity incurs the cost of higher rates of mis-classification; and ii) misclassification varies with thematic scale. We comparied the performance of an algorithm (Chameleon) which operates on interconnectivity and thus is sensitive to the shape and distribution of clusters with that of three traditional algorithms over varying scales.

Results: Chameleon-derived solutions had lower rates of misclassification and only marginally higher heterogeneity than those of k-means in the range 15–60 clusters, but their metrics converged at finer thematic scales. Solutions derived by agglomerative clustering had the best metrics (and divisive clustering the worst) but both produced inferior high-level solutions clusters to those of Chameleon by merging distantly-related clusters.

Conclusions: Our results suggest that Chameleon may have an advantage over traditional algorithms at thematic scales at which data exhibit discontinuities and variable structure, potentially producing more stable solutions (due to lower rates of mis-classification), but scoring lower on traditional metrics of central-tendency. Chameleon's advantages are less obvious in the partitioning of continuous data, however its graph-based partitioning protocol facilitates hierarchical integration of solutions.

Introduction

Vegetation classification is the process of delimiting types of vegetation on the basis of their relative homogeneity and distinctness from other types (van der Maarel & Franklin 2013). Classification facilitates not only the description of vegetation, but also the study of its relationships with the environment and attendant interacting, co-dependant organisms. Vegetation classification is thus a first step to the classification of ecosystems (*sensu* Tansley 1935), and vegetation typologies have come to underpin a wide variety of conservation and natural resource management applications including the selection of protected areas, ecosystem risk assessment and market-based mechanisms such as biodiversity offsets (Bland *et al.* 2019). Despite a relatively short history, the science has spawned a wide range of schools (Whittaker 1978, van der Maarel & Franklin 2013). Increasingly however, vegetation classification centres on the clustering of quantitative plot samples (De Cáceres *et al* ., 2015; 2018). When recorded with systematic procedures, plot samples have the advantage of allowing observations from different sources to be consolidated over time, while computer-generated clustering solutions confer a degree of objectivity in the elucidation of patterns.

The utility of clustering in the development of vegetation classifications is beyond question, although it is complicated by three inter-related problems. First, excepting simulated datasets, there is no agreed external point of reference with which clustering solutions can be compared. Instead, solutions based on field data must be evaluated on internal criteria (Aho *et al* . 2008), either geometric (eg cluster homogeneity) or non-geometric (eg species/cluster fidelity). Since these vary in the way they weight particular characteristics of the solution, the best clustering solution may depend on its application. Second, the hyper-spatial structure of vegetation data is generally unknown. The choice of both clustering algorithm and evaluation metrics therefore requires a user-supplied model which usually (but not invariably (Aho*et al* . 2008)) means clusters are assumed to be spheroidal, if only because the majority of operators default to a few well-tested algorithms (Kent 2011). This is problematic, because algorithms which seek to optimise central tendency can generate sub-optimal solutions when applied to data with irregular structure, and internal metrics which assume a spheroidal model may not be appropriate measures of cluster quality. Third, biases in the both the geographic and environmental distribution of samples means that cluster metrics are often optimised for data which sample the range of floristic variation either unevenly or incompletely. That is, biases may induce irregularities in data structure even if assemblages in the field form a continuum. It is not surprising then, that clustering solutions are notoriously idiosyncratic and highly sensitive to data structure, transformations, choices of algorithm and resemblance measures (Tichy*et al* ., 2014).

The potential limitations of assuming a spheroidal model to data of irregular structure are illustrated in Figure 1. The data are points on a cartesian plain, normally and randomly distributed around each of six pre-defined centroids. The k-means algorithm fails to retrieve the underlying data structure; in i) incorrectly splitting cluster C while merging clusters D and F; and ii) incorrectly splitting clusters C and F to partially merge with clusters A and D, respectively. The resulting solutions appear what Barton *et al* . (2019) termed 'unnatural', although they conceded the vagueness (*sensu* Regan et al, 2002) of the assignation, relying as it does on an appeal to the human eye. Less subjectively, the solution is 'incorrect', for example in Figure 1(ii) in assigning samples that are co-located in space in the region of centroid C to different groups, while drawing in remotely-located samples from the region of centroid A. The implication is there is a high likelihood of alternative solutions arising as further data are added, or if the clustering algorithm is changed or supplied different parameters.

The problem illustrated in Figure 1 arises primarily to the insensitivity of the algorithm to variations in the density of points, however a failure to recover 'natural' or 'correct' clusters of irregular shape has similarly been documented in a wide range of algorithms operating on assumptions of central tendency (Karypis 1999, Zhao & Karypis 2005, Han *et al* . 2012, Barton *et al* . 2019). The core principle underpinning algorithms which seek to retrieve clusters of irregular shape and/or density is sample inter-connectivity. That is, cluster membership depends on interconnections among sample (based on pairwise similarity) rather than shared proximity to an artificial centroid or medoid. Schmidtlein *et al* . (2010), for example, noted two vegetation samples with no species in common could nevertheless share cluster membership provided they were connected in a chain of close neighbours. This implies clusters generated by an algorithm sensitive to irregular data structure are likely to be more heterogeneous than those derived with reference to a spheroidal model, particularly at thematic scale where discontinuities and variation in sample density exist.

Potential irregularities in data structure are rarely accounted for in vegetation classification. Schmidtlein *et al* . (2010) documented a promising approach, however our investigations of their ISOMAP algorithm suggested its "brute force" approach is too computationally demanding for a dataset many thousands of samples (Schmidtlein *et al* . (2010) investigated datasets ranging in size up to 305 samples and warned users of ISOMAP that the algorithm is slow, and not to complain!). Chameleon (Karypis et al., 1999, see methods for a detailed description) is one of several alternative algorithms designed to recover clusters of variable shape which may, therefore, reproduce landscape scale relationships more faithfully than traditional clustering techniques (Han *et al* ., 2012). Chameleon assesses both interconnectivity and closeness of objects as a basis for determining merging decisions, an approach which results in fewer "wrong" decisions than algorithms that consider only one or the other (Karypis*et al* ., 1999). Focussing on interconnectivity allows the algorithm to adapt automatically to the characteristics of the clusters (density and hyperspatial distribution), rather than relying on a static model (eg discrete spherical clusters). Therefore, provided they are strongly interconnected, samples spanning a compositional continuum can be retrieved as a single cluster even if the distribution of samples along the continuum is uneven, because Chameleon is relatively insensitive to variations in hyperspatial density (Han *et al* ., 2012).

We suggest that a failure to take account of the underlying structure of vegetation data is likely to be one factor contributing to idiosyncrasies among clustering solutions, however the affect is likely to be dependent on the expression and nature of discontinuities in the data structure. We postulate that accounting for data structure is more likely to be important at broad thematic scales (as represented by the points in Figure 1 collectively) because discontinuities are likely to arise both naturally (eg between regions which share few species), due to variable data coverage (De Cáceres *et al* ., 2018, Gellie*et al* . 2018) or because environmental gradients are discontinuous in geographic space (Austin 2013). Conversely, there may be no disadvantage in assuming a spheroidal model where clustering essentially amounts to partitioning a continuum (ie partitioning the individual clusters in Figure 1). In this paper, we investigate two hypotheses: i) that an algorithm sensitive to hyperspatial irregularities in the density and arrangement of samples will produce clusters which are likely to be more 'correct' (in the sense that samples are co-located with their close neighbours), but at the cost of poorer internal metrics relative to algorithms that seek to optimise around central tendency; and ii) Differences between the respective algorithms will decline with decreasing thematic scale of cluster solutions. To test these hypotheses, we used a large regional data set of 7541 plot samples to compare the performance of traditional clustering algorithms (k-means, hierarchical agglomerative and divisive) with the Chameleon algorithm using both internal metrics (homogeneity, indicator species) and the concept of 'correctness' which we apply as the misclassification rate: the proportion of samples which do not cluster with their nearest neighbour.

Methods

The Chameleon algorithm

Chameleon models the feature space as a k-nearest neighbour graph (sparse graph) with samples forming vertices connected by links that are proportional to pairwise similarity between samples (Figure 2). The user specifies the number of links between samples (neighbourhood range) and then in the first phase, links are progressively dissolved (in order of increasing similarity) until a user-specified number of sub-partitions has formed. In this partitioning phase the algorithm seeks to minimise the summed length of all links hence minimising the affinity between samples in different sub-partitions (Karypis, 1999). Sub-partitions are then (optionally) merged using a hierarchical agglomerative clustering algorithm to resolve the number of groups required of the solution. An advantage of this approach is that it encapsulates the concept of environmental /compositional continua by weighting cluster interconnectivity over homogeneity. That is, samples that are distantly related may still share a cluster if they are linked by strongly interconnected neighbours. One of the key features of the Chameleon algorithm is that the structure of inter-sample relationships is preserved through the partitioning phase because co-membership of sub-partitions is dependent on pairwise inter-sample connectivity. In contrast, traditional algorithms merge or split samples progressively and the outcome at each step depends on comparing samples with intermediate clusters, the compositional characteristics of which are artificial and reflect the range of the samples merged (Han *et al* ., 2012).

Study Area

The study area encompassed the South East Highlands and Australian Alps Bioregions of the state of NSW, Australia (Thackway & Creswell, 1995), an area of 96,089 km^2 encompassing mountains and plateaus of the Great Dividing Range. Average annual precipitation ranges from 460 —- 2,344 mm and mean annual temperatures are 3 – 16° C. The area is underlain by a complex series of heavily folded metamorphosed sedimentary rocks deposited from the Ordovician to Devonian periods and interspersed with numerous granite intrusions and, to a much lesser extent, basalts deposited in the Paleogene.

Primary factors influencing the distribution of vegetation formations in our study area include temperature, rainfall, topography, soils and drainage (Jenny, 1983; Costin, 1954; Beadle, 1981; Keith, 2004). Alpine assemblages are restricted to elevations more than 1830 m above sea level where winter temperature minima fall below the physiological tolerance of trees (Keith, 2004). Tree cover progressively increases with decreasing elevations severity of winter conditions declines. Sub-alpine grassy woodlands occupy the sub-alpine tracts, characteristically with short gnarled trees and a large compliment of cold-tolerant species also found in the alpine zone. On the southwest flank of the Great Divide, sub-alpine woodlands grade into tall wet sclerophyll forests, sustained by high orographic rainfall originating in south-westerly air flows. To the east, depending on soils lithology texture and fertility, sub-alpine woodlands grade into either Dry Sclerophyll Forest or Grassy Woodlands as annual rainfall declines in the shadow of the Divide. Grasslands replace Grassy Woodlands in frost hollows, the heaviest-texture soils and the most moisture-limited sites (Costin, 1954). Further east of the tablelands, grasslands and grassy woodlands are replaced by mosaics of wet and dry sclerophyll forests on the escarpment ranges as rainfall increases with increasing elevation and exposure to oceanic weather systems (Keith, 2004). Wetlands occur throughout the bioregions in areas of impeded drainage, while heathlands are among the local expressions of edaphic and topographic variation.

Compilation of Floristic Data

We sourced a total of 7541 floristic plot samples from a database compiled and administered by the Department of Planning, Industry and Environment (DPIE, 2019). These comprised all survey data collected in (or within 25 km of) the South East Highlands and Australian Alps Bioregions which met the following criteria: i) the sample location was recorded with an accuracy of < c.100m; ii) the sample area was 0.04 ha; and iii) all vascular plant species were recorded. Individual species records were reviewed and modified to resolve inconsistencies in taxonomy (see Methods in Tozer *et al* ., 2010). Taxa identified only at the generic level were removed along with records of naturalised species. Cover-abundance scores were transformed to presence-absence to eliminate the possible effects of bias in cover-abundance estimates by different observers. This transformation was considered an appropriate strategy to achieve a balance between information-loss and maximising the pool of available data in circumstances where the data set is both large and likely to be heterogeneous (Goodall, 1978).

Chameleon performance evaluation

We performed all Chameleon analyses on pairwise Bray-Curtis compositional similarities between samples (Clarke, 1993) using the scluster function in CLUTO software version 2.1.2 (Karypis, 1999). First, since we found little information in the literature to guide parameter-setting, we assessed solutions of 15 clusters over a range of neighbourhood sizes (15 – 1000 neighbours), degrees of sub-partitioning (up to 500 sub-partitions or agglomerative phase omitted) and linkage functions (single or complete) (Table 1). We focused carried out our initial trials using the cluster-weighted single-link criterion function in the agglomerative phase, as recommended for non-spherical clusters (Karypis, 1999). For each solution we calculated average pairwise within-cluster association (homogeneity) and the proportion of samples located in clusters other than that of their nearest neighbour (misclassification rate). We found using the single linkage function caused chaining (*sensu* Peet & Roberts 2013) when the number of sub-partitions specified was larger than 30. We repeated the relevant trials using an option forcing Chameleon to prioritise large clusters over small in the partitioning phase. On the basis of the preliminary results, we undertook subsequent analyses using the complete linkage function and assessed performance over a range of thematic scales (15 – 250 clusters) and degrees of sub-partitioning (30 – 500 sub-partitions) with neighbourhood size fixed at either 30 or 1000 (Table 1).

Comparison of algorithms

We compared Chameleon cluster member-sets with those derived using: i) k-means clustering; ii) flexible unweighted pair-group averaging with arithmetic mean (Belbin *et al* . 1992); and iii) polythetic-division (MacNaughton-Smith *et al* ., 1965; Belbin*et al* ., 1984). We transformed the adjacency matrix supplied to scluster to dissimilarity (1-simmilarity) and used each algorithm to compute solutions ranging from 15 – 250 clusters (Table 1.) We characterised each solution in terms of homogeneity and misclassification rate (as described above), the number of species occurring at higher frequencies within each cluster than in the dataset as a whole (cumulative hypergeometric probability >0.999) and the number of species with standardised phi > 0.35 (Tichy & Chytry 2006).

Comparing clustering solutions with a reference classification

We assessed the extent to which clustering solutions (15 classes) produced by each algorithm retrieved species-sets characterising the units of an established subcontinental-scale vegetation classification that covers 800,000 km^2 in southeastern Australia (Keith 2004), including the study area (c. 11% of total area). The reference classification was developed from the top-down based on an extensive review of vegetation studies, field reconnaissance and qualitative synthesis of vegetation maps available at the time (Keith 2004). Its highest level of classification (vegetation formation) is based on structural/physiognomic features. Formations are subdivided into vegetation classes based on geographically distinct expressions of structural and compositional features. Fifteen of 99 vegetation classes recognised in the reference classification are mapped within the study area and are described with lists of indicative species (Keith, 2004). For each clustering solution, we identified the species diagnostic of each cluster as those with a frequency of occurrence statistically higher within the cluster samples than across the dataset as a whole (cumulative hypergeometric probability > 0.999). We compared these with the species identified as diagnostic of the reference classes, compiling a confusion matrix with the units of the respective classifications as rows and columns and cell values calculated as the proportion of reference class species that were identified as diagnostic of each cluster class.

Results

A summary of the analytical trials performed and a brief synopsis of the results is contained in Table 1. A detailed description of the results follows:

Performance of Chameleon under combinations of varying parameters- single linkage

Trends in the mis-classification rate and average within cluster homogeneity in Chameleon cluster solutions generated using the weighted single-link functions are summarised in Figure 3. The misclassification rate rose with increasing neighbourhood size (Figure 3A). This result may reflect aberrations caused by forcing members of small clusters to forge links with samples in other clusters as illustrated by Chameleon's attribution of the simulated data we presented in Figure 1 given neighbourhoods of different sizes (Figure 4). Solutions derived by agglomeration from 30 sub-partitions had consistently lower rates of misclassification, but beyond 30 sub-partitions solutions became increasing uneven (chaining) and mis-classification rates became meaningless because a high proportion of samples were concentrated in few clusters. The problem of chaining was not corrected by directing the algorithm to prioritise large clusters over small in the partitioning phase however more even clusters were produced when the cluster-weighted complete link function was employed in the agglomerative phase of the algorithm and subsequent analyses were performed using this option, as described in the next section. There was no clear trend in within-cluster homogeneity with increasing neighbourhood size when the agglomeration phase was omitted (Figure 3B). Solutions derived by agglomeration from 30 sub-partitions had highest homogeneity with a neighbourhood size of 100. Beyond 30 sub-partitions the data showed no clear trend and varied erratically depending on the uneven-ness of the solutions.

Clusters of 15 solutions generated using the cluster-weighted complete link function exhibited higher rates of mis-classification and lower within-cluster homogeneity when either neighbourhood size (n) or the number of sub-partitions (a) in the agglomerative phase were increased, although increasing n disproportionately affected the mis-classification rate while increasing a disproportionately affected cluster homogeneity (Figure 5).

Both the rate of mis-classification and within-cluster homogeneity increased with increasing thematic resolution (Figure 6). Chameleon solutions derived using small neighbour sizes and either: modest numbers of sub-partitions (twice the number of classes in the solution); or with the agglomeration phase omitted, were better (lower rates of misclassification and higher homogeneity) than those derived with the divisive algorithm, but worse than those derived with the agglomerative algorithm (Figure 6). However, 15- class solutions derived by Chameleon were more even than those produced by either the agglomerative or divisive algorithms (Figure 7). Chameleon solutions were better than those of k-means at broad thematic scales (15 – 60 classes) but equivalent at finer scales (90 – 250 classes). Chameleon produced more even 15-class solutions than k- means (Figure 7).

Clusters derived by Chameleon solutions were generally characterised by fewer diagnostic species than those derived using the traditional algorithms (Table 2), however species diagnostic of Chameleon clusters corresponded more with those characterising units of a reference classification for our study area than those diagnostic of cluster derived by agglomerative or divisive algorithms, both in the range of units represented and with less overlap between unrelated units (Table 3a, 3b, 3c). Clusters derived by k-means retrieved units of the reference classification with efficiency similar to Chameleon (Table 3d).

Discussion

Is there evidence of irregular structure in vegetation data reflected in the performance of different algorithms?

We hypothesised that in cases where the structure of vegetation data is variable (irregular shaped clusters or variable density), an algorithm sensitive to such variability would perform better (lower rates of mis-classification) than one that optimises central-tendency (more homogenous clusters). While the structure of our vegetation data is unknown, it is unlikely to be regular, neither continuous along environmental gradients nor arranged in discrete clusters. Theory and empirical evidence suggest that assemblages of species form multi-dimensional continua (Whittaker, 1975; Goodall, 1978; Kent, 2011). However, discontinuities may arise where environmental gradients are either discontinuous in geographic space, or parts of the environmental spectrum are not represented (Austin, 2013). Discontinuities are also likely to arise in our data at broader thematic scales due to biases in the distribution in sample (Gellie *et al* . 2018) and they patently exist at continental scales between climatically similar sub-continental regions which are separated by water or large areas with unsuitable climate and so share few species (Tozer *et al* . 2017). We further hypothesised, therefore, that Chameleon's primary advantage was likely to be in the elucidation of upper-hierarchal clusters.

Overall, the results of our analyses support both hypotheses, although it is clear that: i) the utility of the different clustering methods cannot be encapsulated solely in terms of cluster homogeneity and rates of misclassification, ii) internal evaluators can be misleading in terms of cluster quality; and iii) the superior performance of Chameleon in elucidating upper-hierarchical clusters is entirely dependent on selecting appropriate parameters from an infinite range of combinations. The clearest support for our hypotheses was evident in the comparison between solutions derived using Chameleon clusters with those derived by k-means over the range from 15 – 250 clusters. Chameleon's best 15 and 30 cluster solutions exhibited significantly lower rates of mis-classification than those of k-means at the cost of an increase in heterogeneity (Figure C), while at progressively higher levels of thematic detail (60 – 250 clusters) there was a convergence in the respective metrics. We speculate that increasing rates of misclassification at finer thematic scale is indicative of the partitioning of a continuum. That is, at fine thematic scales communities increasingly intergrade such that the proportion of their (ever decreasing) member-sets which most closely resemble samples in adjacent clusters increases. If there was indeed variability in the structure of our data at broad thematic scales, then the algorithms performed as hypothesised. We conclude there was a clear advantage in using Chameleon over k-means to elucidate our upper-hierarchical clusters (and relatively little cost), but no apparent advantage at finer thematic scales in terms of cluster metrics. However, since Chameleon solutions of progressive finer scale can be produced by continually partitioning the sparse graph, the algorithm potential offers a straightforward method of integrating plot-based classifications at multiple hierarchical scales.

Accounting for the performance of agglomerative and divisive clustering algorithms is more complicated. First, on the basis of cluster homogeneity and rates of mis-classification, our agglomerative algorithm performed better than either Chameleon or k-means, scoring higher on both metrics at all levels of thematic detail, while our divisive algorithm scored worse (Figure C). Both, however, produced 15-cluster solutions of much greater unevenness in membership numbers than k-means or Chameleon (Figure D) which, if evidence of chaining (*sensu*Peet & Roberts 2013), could suggest that both solutions were less informative in relation to the nature of upper-hierarchical groupings. Conversely, our three traditional algorithms scored equally highly in terms of the number of diagnostic species and clearly higher than the best Chameleon solutions, suggesting that unevenness in cluster membership numbers could, in fact, be symptomatic of biases in the distribution of samples among 'natural' clusters, and that the three traditional algorithms performed better in detecting these uneven clusters (as evidenced by higher numbers of diagnostic species).

Comparisons with a reference classification suggest unevenness in the cluster size is more likely to be indicative of chaining, because indicator species of the largest clusters tended to represent large numbers of known classes, some of which are relatively distantly related, a phenomenon most strongly evident in the agglomerative and divisive solutions (Tables 3a-c). This reflects a well-known weakness of agglomerative or divisive methods which incorporate merge or split decisions based on the aggregate properties of clusters. Such methods require either unrealistic assumptions concerning the structure of the data and/or sequential merge/split decisions which cannot be reversed, and which are necessarily sensitive to the composition of the dataset (Han *et al* ., 2012). While we did not evaluate the quality of solutions of greater than 15 classes, we suggest our agglomerative algorithm outperformed all others in producing 250-cluster solutions with low rates of misclassification and high homogeneity, but that subsequent, upper-hierarchical groupings because progressively less meaningful because of poor merging decisions. We conclude that Chameleon and k-means generated the most informative solutions of 15 clusters with the former perhaps better representing the natural structure of the data while the latte produced more homogeneous groupings.

Are 'natural' clusters necessarily less homogeneous?

Although a degradation of cluster homogeneity is implicit in our model, the degree to which this is realised is likely to be highly dependent on the structure of individual datasets. In our case study, the mis-classification rate achieved by Chameleon was half that of k-means at the cost of a 10% reduction in cluster homogeneity. We speculate that if the clusters Chameleon retrieved in our dataset are indeed irregular shapes, then our results suggest they are unlikely to be highly elongated, and variability in our data structure tends toward uneven density rather than irregular shape.

The question of whether 'natural' clusters necessarily have fewer diagnostic species is more difficult to resolve based on our analyses.*A priori* , we inclined to the notion that more heterogenous clusters would mean fewer diagnostic species, the pattern reflected in our results, however Schmidtlein et al. (2010) demonstrated that Isopam, an algorithm that adapts to irregular cluster shapes, consistently out-performed other algorithms in terms of the number of indicator species (*sensu* Dufrêne & Legendre 1997) and was also highly ranked in terms of the number of species with standardized phi >0.35 (Tichy & Chytry 2006). Higher numbers of diagnostic species could reflect the sampling of a wider species pool, since samples sharing no species can occupy the same cluster if comprise an interconnected neighbourhood (Schmidtlein et al. 2010). However, it is not clear that higher numbers of diagnostic species is not an artefact of Isopam's partitioning of the ordinations space by medoids, notwithstanding the fact the ordination axes are adjusted to accommodate non-linearities (and hence irregularities).

On the evidence of our results, we conclude that our original contention is supported, that cluster solution derived by algorithms sensitive to data structure are unlikely to be as compact or homogenous as those derived by optimising central tendency, although the differences may not always be pronounced, depending on the characteristics of individual datasets and thematic scale of investigation. In that case, we suggest that further research is required into metrics which give insights into how well cluster solutions model the structure of vegetation data (eg within-cluster inter-connectedness, mis-classification rates) to better understand the potential trade-offs involved in maximising homogeneity or indicator values.

Are natural clusters likely to be more stable/robust to new data?

Clustering solutions are notoriously sensitive to classification protocols, and it has generally proven difficult to retrieve the units of individual CCSs via meta-analysis of combined data (Tichý *et al* . 2011, 2014). Wiser & De Cáceres (2013) and Tichý *et al* . (2014) characterised this problem in terms of the need to preserve units of one or more CCSs while allowing for previously unrecognised units to be identified following the acquisition of new plot data. Their respective solutions comprise alternative forms of semi-supervised clustering, promising approaches that allow for units to be "fixed" by specifying their plot membership a priori while allowing for unattributed plots to form new clusters. The question of when units should be "fixed" must still be addressed. If the problem arises either because algorithms cluster irregular data in idiosyncratic ways or there are biases in the distribution of samples in compositional then some understanding of the underlying data structure is likely to be informative.

In theory, algorithms sensitive to data structure may reduce the extent of this problem, at least at some thematic scales. Tozer et al. (2022) concluded Chameleon's novel approach to modelling inter-sample relationships greatly facilitated the revision of an earlier broad-thematic classification of forested wetlands based on substantially fewer plot samples (Keith & Scott 2005). Unlike many traditional methods which incorporate merge or split decisions based on the aggregate properties of clusters, Chameleon operates on inter-connected neighbourhood sets structured, in Tozer *et al* 's (2022) case, on the same similarity metric used in the original analysis. They considered these features pivotal, because the algorithm could potentially minimise the impact of incremental additions of new data by retaining connections between samples from the original set (Tozer *et al* . 2022) (although they speculated that this could best achieved by specifying a large neighbourhood which, on the basis of our results, we suggest is not appropriate for minimising the rate of misclassification). Tozer *et al* . (2022) reasoned that since Chameleon dissolves connections between relatively weakly-connected samples in the partitioning phases, strong pairwise relationships between samples underpinning clusters in the original analysis were potentially preserved (and reflected more faithfully) in their new Chameleon-derived clusters (provided they were not displaced by a sufficiently large number of more strongly inter-connected samples). This interesting feature requires further study.

Conclusion

Our study demonstrates that scale-dependent irregularities in vegetation data can exist and potentially affect the utility and stability of clustering solutions underlying vegetation classification schemes. The existence of clusters of irregular shape and density implies that novel metrics are required in their evaluation because clusters which are 'natural' in the sense they reflect human responses to visual cues (Barton *et al* . 2019) are unlikely to score well on traditional metrics that assume a spheroidal model (Aho *et al* . 2008). Evaluating the utility of such cluster solutions requires metrics which assess inter-connectivity rather than central tendency.

While the results presented here demonstrate the potential utility of the Chameleon algorithm in vegetation science, there exist several issues requiring further investigation. Although Chameleon produced informative solutions at broad thematic scales, solutions derived with different parameters varied markedly, and some were clearly inferior to those of the traditional algorithms we evaluated. While its theoretical advantages are widely cited, we found very few examples in the literature to guide how Chameleon should be parameterised, and none pertaining to analysis of vegetation data. Although the algorithm can be implemented on wide range of distance metrics, we opted to import a distance matrix which underpins Consistent Classification Sections (CCS,*sensu* De Cáceres *et al* . 2015) within our study area in order to maximise the potential for integrating our results with those CCSs. On the basis of our trials, we recommend the use of small neighbourhood sizes over large and either omitting the agglomerative phase or restricting the number of partitions to no more than twice the number of samples in the desired solution. Although Karypis (2003) recommended using a cluster-weighted single linkage function in the implementation of Chameleon, we found this induced chaining in our solutions, while the cluster-weighted complete linkage function produced satisfactory results and we recommend this function if an agglomerative step is employed. Further experimentation with each of these parameters using other datasets is clearly required.

Finally, there is some uncertainty in relation to how the algorithm can be implemented. We employed the Cluto clustering package (Karypis 2003) distributed by Chameleon's authors, however we noted some inconsistencies in relation to the parameters offered compared to the description of the algorthim (Karypis *et al* . 1999). Furthermore, Barton *et al.* (2019) have suggested Cluto's implementation does not entirely embody the Chameleon concept. Barton *et al.* (2019) offer an alternative implementation which deserves evaluation, although it relies on an different partitioning algorithm because the original is proprietary protected.

In summary, while our results support the notion the Chameleon algorithm is theoretically better suited to the task of elucidating vegetation classes, the characteristics of its solutions, and the ways in which these improve upon those retrieved by traditional clustering approaches requires further quantification. We suggest this is a worthwhile endeavour because Chameleon offers a conceptually simple model, can process very large datasets quickly and potentially presents a solution to the problem of integrating plot-based classifications across hierarchical levels.

Data availability statement

CLUTO software modules are available for download from Karypis Lab website (http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download). Plot data used in our analyses are available at: https://www.environment.nsw.gov.au/research/Vegetationinformationsystem.htm (NSW DPIE 2020, accessed 2nd August 2016). All analyses were performed on an adjacency matrix of similarity (1-Bray-Curtis dissimilarity) between the objects to be clustered. Data were imported in a plain text file with $n + 1$ lines, the first line containing the number of rows and the remaining n lines containing adjacency values for each row (Karypis 2003).

References

Aho, K., Roberts, D.W. and Weaver, TR. (2008) Using geometric and non-geometric internal evaluators to compare eight vegetation classification methods. *Journal of Vegetation Science* 19(4), 549-562.

Austin, M.P. (2013) Vegetation and Environment: Discontinuities and Continuities. In: E. van de Maarel and Franklin, J. (Eds) *Vegetation Ecology* , Second Edition. John Wiley & Sons, Ltd, pp. 71-106.

Barton, T., T. Bruna and P. Kordik (2019) Chameleon 2: an improved graph-based clustering algorithm. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13(1), 1-27.

Beadle, N.C.W. (1981) *The vegetation of Australia* . Cambridge: Cambridge University Press.

Belbin, L., Faith, D.P. and Minchin P.R. (1984) Some algorithms contained in the Numerical Taxonomy Package NTP. Technical Memorandum 84/23.

Belbin, L., Faith, D.P. and Milligan, G.W. (1992) A comparison of two approaches to ss-flexible clustering. *Multivariate Behavioural Research* , 27, 417-433.

Bland, L.M., Nicholson, E., Miller, R.M., Andrade, A., Carre, A., Etter, A., Ferrer-Paris, J.R. Herrera, B., Kontula, T., Lindgaard, A., Pliscoff, P., Skowno, A., Valderrabano, M., Zager, I. & Keith, D.A. (2019) Impacts of the IUCN Red List of Ecosystems on conservation policy and practice. *Conservation Letters* **e12666** [doi: 10.1111/conl.12666].

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* **18** (1), 117-143.

Costin, A.B. (1954) *A study of the ecosystems of the Monaro region of New South Wales, with special reference to soil erosion* . AH Pettifer, Government printer.

De Caceres, M., Chytry, M., Agrillo, E., Attorre, F., Botta-Dukat, Z., Capelo, J., Czucz, B., Dengler, J., Ewald, J. and Faber-Langendoen D. (2015) A comparative framework for broad-scale plot-based vegetation classification. *Applied Vegetation Science,* 18(4), 543-560.

De Caceres, M., Franklin, S.B., Hunter, J.T., Landucci, F., Dengler, J., & Roberts, D.W. (2018). Global overview of plot-based vegetation classification approaches. *Phytocoenologia* , *48* (2), 101-112.

DPIE (2019) Department of Planning Infrastructure and Environment, NSW Government - *BioNet Systematic Flora Survey* . Available at https://www.environment.nsw.gov.au/research/VISplot.htm [Accessed 14 November 2020]

Dufrene, M. and Legendre, P. (1997) Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological Monographs* 67, 345-366.

Goodall, D.W. (1978) Numerical methods of classification. In Whittaker, R.H. (Ed), *Classification of plant communities* . Springer pp. 247-286.

Han, J., Kamber, M. and Pei, J. (2012) *Data Mining: Concepts and Techniques,* Third Edition. Waltham, Massachusetts: Morgan Kaufmann Publishers.

Gellie, N., Hunter, J.T., Benson, J.S., Kirkpatrick, J.B., Cheal, D.C., McCreery, K. and Brocklehurst, P. (2017) Overview of plot-based vegetation classification approaches within Australia.*Phytocoenologia* , 48(2), 251-272.

Jenny, H. (1983). *Factors of soil formation - a system of quantitative pedology* . New York: McGraw-Hill.

Karypis, G., Han, E.H. and Kumar, V. (1999) Chameleon: Hierarchical clustering using dynamic modelling. *Computer* , 32(8), 68-75.

Karypis, G. (2003) *CLUTO A Clustering Toolkit* , Technical Report: #02-017. University of Minnesota, Department of Computer Science Minneapolis, MN 55455.

Keith, D.A. (2004) *Ocean shores to desert dunes: the native vegetation of NSW and the ACT* . Hurstville: New South Wales Department of Environment and Conservation.

Kent, M. (2011). *Vegetation description and data analysis: a practical approach* . Chichester: John Wiley & Sons.

MacNaughton-Smith, P., Williams, W.T. and Dale, M.B. (1964) Dissimilarity analysis: a new technique of hierarchical sub-division.*Nature* , 202, 1034-1035.

Peet, R.K. and Roberts, D.W. (2013) Classification of natural and semi-natural vegetation. In: E. van de Maarel and Franklin, J. (Eds) *Vegetation Ecology* , Second Edition. John Wiley & Sons, Ltd, pp. 28-70.

Regan, H.M., Colyvan, M. and Burgman, M.A. (2002) A taxonomy and treatment of uncertainty for ecology and conservation biology.*Ecological applications* 12(2), 618-628.

Schmidtlein, S., Tichý, L. Feilhauer, H. and Faude U. (2010) A brute-force approach to vegetation classification. *Journal of Vegetation Science* 21(6), 1162-1171.

Tansley, A.G. (1935) The use and abuse of vegetational concepts and terms. *Ecology* , **16** (3), 284-307.

Thackway, R. and Creswell, I.D. (1995) *An interim biogeographic regionalisation of Australia: a framework for establishing the national system of reserves* , Version 4. Canberra: Australian Government.

Tichy , L., and Chytry , M. (2006) Statistical determination of diagnostic species for site groups of unequal size. *Journal of Vegetation science* , *17* (6), 809-818.

Tichy, L., Chytry, M. and Botta-Dukat, Z. (2014) Semi-supervised classification of vegetation: preserving the good old units and searching for new ones. *Journal of Vegetation Science* 25, 1504–1512.

Tozer, M.G., Turner, K., Simpson, C., Keith, D.A., Beukers, P., MacKenzie, B., Tindall, D. and Pennay, C. (2010) Native vegetation of southeast NSW: a revised classification and map for the coast and eastern tablelands. *Cunninghamia* 11(3), 359-406.

Tozer, M.G., Simpson, C.C. and Jansens, I.B. (2017). Biogeography of Australia's dry sclerophyll forests: drought, nutrients and fire. In: Keith, D.A. (ed.) *Australian Vegetation* (3rd edition). Cambridge University Press, Cambridge.

Tozer, M. G., Simpson, C. S., & Keith, D. A. (2022). Subtropical-temperate forested wetlands of coastal south-eastern Australia–an analysis of vegetation data to support ecosystem risk assessment at regional, national and global scales. *Pacific Conservation Biology* doi:10.1071/PC21028.

Van Der Maarel, E. and Franklin, J. (2013) Vegetation ecology: historical notes and outline. In: E. van de Maarel and Franklin, J. (Eds) *Vegetation Ecology* , Second Edition. John Wiley & Sons, Ltd, pp. 1-27.

Whittaker, R.H. (1975) *Communities and Ecosystems* . New York: Macmillan.

Wiser, S.K. and Cáceres, M. (2013) Updating vegetation classifications: an example with New Zealand's woody vegetation. *Journal of Vegetation Science* , 24(1), 80-93.

Zhao, Y. and Karypis, G. (2005) Data clustering in life sciences.*Molecular biotechnology* , 31(1), 55-80.

Tables

Table 1: Summary of analytical trials undertaken, their purpose and results. Colours in column match indicative results plotted in Figures 3-6.

Algorithm	Number of clusters	Nei
Chameleon	15	15 -
Chameleon	15	15 -
Chameleon	15	30

Algorithm	Number of clusters	Nei
Chameleon	15, 30, 60, 90, 120, 150 200, 250	30
Chameleon	15, 30, 60, 90, 120, 150 200, 250	30
Chameleon	15, 30, 60, 90, 120, 150 200, 250	100
k-means	15, 30, 60, 90, 120, 150 200, 250	NA
flexible unweighted pair-group averaging with arithmetic mean (Belbin *et al.* 1992)	15, 30, 60, 90, 120, 150 200, 250	NA
polythetic-division (MacNaughton-Smith *et al.*, 1965; Belbin *et al.*, 1984)	15, 30, 60, 90, 120, 150 200, 250	NA

Table 7: Total number of diagnostic species across all classes as determined by frequency (statistically higher than background frequency) or standardised phi (Tichy & Chytry 2006).

Algorithm	total number of species with class hypergeometric probability > 0.999	total number of spe
k-means	3615	303
Agglomerative	3478	282
Divisive	3118	278
Chameleon (n=10, a = 15)	3569	252
Chameleon (n=40, a = 15)	3572	268
Chameleon (n=100, a = 15)	3416	269
Chameleon (n=1000, a = 15)	3416	269
Chameleon (n=1000, a = 60)	4646	549

Table 3a:: Proportion of characteristic species for each reference class (rows) shared with clusters from the Chameleon algorithm (15 clusters based on neighbourhood range of 1000 samples agglomerated from 30 sub-partitions). Dark grey indicates proportions>0.7, pale grey proportions >0.5. Cells with the same shading in column one comprise members of the same formation.

Cluster (15)	6	7	9	0	4	5	8	14	3	15	1	12	2
Alpine Herbfields	0.88	0.00	0.00	0.12	0.19	0.19	0.19	0.00	0.00	0.00	0.00	0.19	0.08
Alpine Bogs and Fens	1.00	0.00	0.00	0.04	0.04	0.07	0.04	0.11	0.04	0.00	0.00	0.21	0.04
Alpine Heaths	0.93	0.00	0.07	0.07	0.07	0.00	0.00	0.04	0.04	0.00	0.00	0.37	0.04
Alpine Fjaeldmarks	1.00	0.00	0.00	0.00	0.00	0.06	0.00	0.06	0.06	0.00	0.00	0.06	0.00
Southern Tablelands DSF	0.03	0.80	0.14	0.40	0.43	0.31	0.09	0.06	0.06	0.09	0.23	0.00	0.26
Southern Escarpment WSF	0.00	0.03	0.88	0.38	0.16	0.00	0.00	0.00	0.00	0.16	0.00	0.03	0.16
Montane Wet Sclerophyll Forests	0.11	0.09	0.49	0.66	0.20	0.06	0.06	0.06	0.06	0.14	0.06	0.14	0.11
Southern Tableland WSF	0.02	0.17	0.41	0.68	0.66	0.27	0.15	0.00	0.02	0.12	0.05	0.07	0.44
Subalpine woodlands	0.19	0.24	0.19	0.62	0.57	0.22	0.14	0.03	0.05	0.08	0.08	0.11	0.30
Tableland Clay GW	0.08	0.05	0.14	0.41	0.65	0.46	0.46	0.00	0.03	0.03	0.00	0.14	0.38
Southern Tablelands GW	0.00	0.26	0.14	0.42	0.79	0.74	0.49	0.02	0.00	0.05	0.02	0.02	0.47
Temperate_Montane_grasslands	0.07	0.04	0.00	0.19	0.48	0.70	0.85	0.00	0.00	0.00	0.00	0.22	0.44
Southern Montane Heaths	0.03	0.24	0.07	0.10	0.14	0.07	0.03	0.59	0.38	0.21	0.28	0.07	0.03
Sydney Montane Heaths	0.00	0.08	0.00	0.04	0.04	0.04	0.00	0.42	0.92	0.50	0.25	0.00	0.00
Sydney Montane DSF	0.00	0.20	0.10	0.07	0.10	0.00	0.00	0.13	0.63	0.80	0.43	0.00	0.10
South East DSF	0.04	0.37	0.33	0.11	0.11	0.02	0.00	0.20	0.43	0.65	0.50	0.00	0.17
Montane Bogs and Fens	0.49	0.04	0.23	0.05	0.15	0.04	0.09	0.32	0.21	0.06	0.02	0.72	0.04
Montane Lakes	0.23	0.00	0.00	0.05	0.09	0.05	0.09	0.00	0.00	0.00	0.00	1.00	0.00
Central Gorge DSF	0.00	0.20	0.04	0.15	0.14	0.16	0.11	0.20	0.02	0.18	0.23	0.00	0.75

Table 3b: Proportion of species characteristic of each structural/physiognomic class that are diagnostic of units of a 15-cluster solution derived using polythetic division.). Dark grey indicates proportions>0.7, pale grey proportions >0.5. Cells with the same shading in column one comprise members of the same formation.

Cluster	2	1	11	8	7	5	9	3	4	6	10	12	13
Alpine Herbfields	0.73	0.00	0.00	0.15	0.23	0.00	0.00	0.35	0.27	0.00	0.00	0.04	0.00
Alpine Bogs and Fens	0.57	0.00	0.00	0.07	0.04	0.07	0.00	0.79	0.25	0.00	0.00	0.00	0.00
Alpine Heaths	0.89	0.00	0.07	0.26	0.00	0.04	0.00	0.26	0.11	0.04	0.04	0.07	0.04
Alpine Fjaeldmarks	0.88	0.00	0.00	0.00	0.06	0.06	0.00	0.31	0.06	0.00	0.00	0.06	0.00
Southern Tablelands DSF	0.00	0.80	0.14	0.43	0.26	0.00	0.03	0.00	0.00	0.14	0.03	0.03	0.03
Southern Escarpment WSF	0.03	0.06	0.91	0.59	0.00	0.00	0.44	0.00	0.03	0.41	0.03	0.13	0.22
Montane Wet Sclerophyll Forests	0.14	0.14	0.40	0.83	0.06	0.03	0.14	0.00	0.09	0.20	0.03	0.03	0.11
Southern Tableland WSF	0.02	0.20	0.34	0.83	0.44	0.00	0.24	0.00	0.15	0.54	0.07	0.17	0.29
Subalpine woodlands	0.22	0.24	0.14	0.78	0.24	0.00	0.05	0.00	0.08	0.27	0.05	0.08	0.11
Tableland Clay GW	0.08	0.08	0.14	0.57	0.59	0.03	0.08	0.03	0.14	0.27	0.00	0.05	0.16
Southern Tablelands GW	0.00	0.21	0.12	0.49	0.79	0.00	0.12	0.00	0.05	0.28	0.05	0.09	0.21
Temperate_Montane_grasslands	0.04	0.04	0.00	0.30	0.89	0.00	0.00	0.04	0.19	0.11	0.00	0.00	0.00
Southern Montane Heaths	0.00	0.76	0.00	0.14	0.07	0.21	0.07	0.00	0.03	0.03	0.00	0.00	0.03
Sydney Montane Heaths	0.00	0.67	0.04	0.04	0.04	0.92	0.00	0.04	0.00	0.04	0.04	0.04	0.04
Sydney Montane DSF	0.00	0.97	0.07	0.10	0.03	0.33	0.03	0.00	0.00	0.10	0.00	0.03	0.07
South East DSF	0.02	0.85	0.28	0.13	0.07	0.26	0.15	0.00	0.00	0.24	0.02	0.09	0.20
Montane Bogs and Fens	0.13	0.04	0.00	0.32	0.13	0.19	0.00	0.64	0.74	0.06	0.00	0.04	0.00
Montane Lakes	0.05	0.00	0.00	0.05	0.14	0.00	0.00	0.14	1.00	0.00	0.00	0.05	0.00
Central Gorge DSF	0.00	0.27	0.07	0.07	0.59	0.00	0.50	0.00	0.00	0.43	0.05	0.09	0.39

Table 3c: Proportion of species characteristic of each structural/physiognomic class that are diagnostic of units of a 15-cluster solution derived using pairwise unweighted group-averaging.). Dark grey indicates proportions>0.7, pale grey proportions >0.5. Cells with the same shading in column one comprise members of the same formation.

Cluster	14	13	15	7	4	5	11	1	2	8	10	6	3
Alpine Herbfields	0.58	0.46	0.46	0.12	0.00	0.15	0.00	0.19	0.15	0.04	0.00	0.00	0.04
Alpine Bogs and Fens	0.50	0.79	0.43	0.07	0.00	0.18	0.00	0.07	0.00	0.00	0.11	0.00	0.04
Alpine Heaths	0.85	0.30	0.56	0.07	0.07	0.41	0.07	0.07	0.00	0.04	0.07	0.00	0.00
Alpine Fjaeldmarks	0.25	0.31	0.88	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.06	0.00	0.00
Southern Tablelands DSF	0.40	0.49	0.03	0.74	0.23	0.40	0.63	0.31	0.14	0.34	0.43	0.43	0.29
Southern Escarpment WSF	0.22	0.34	0.06	0.22	0.97	0.44	0.34	0.13	0.00	0.09	0.25	0.06	0.09
Montane Wet Sclerophyll Forests	0.34	0.37	0.17	0.31	0.54	0.80	0.43	0.23	0.06	0.20	0.29	0.09	0.03
Southern Tableland WSF	0.46	0.71	0.05	0.39	0.49	0.66	0.68	0.66	0.17	0.24	0.44	0.20	0.32
Subalpine woodlands	0.70	0.54	0.14	0.43	0.32	0.78	0.51	0.46	0.14	0.32	0.43	0.19	0.19
Tableland Clay GW	0.35	0.46	0.05	0.22	0.24	0.43	0.38	0.68	0.43	0.14	0.27	0.08	0.38
Southern Tablelands GW	0.51	0.53	0.02	0.44	0.23	0.33	0.51	0.72	0.56	0.21	0.33	0.14	0.42
Temperate_Montane_grasslands	0.33	0.48	0.00	0.19	0.04	0.22	0.26	0.59	0.78	0.11	0.19	0.04	0.33
Southern Montane Heaths	0.28	0.17	0.03	0.38	0.14	0.10	0.45	0.10	0.03	0.66	0.38	0.34	0.10
Sydney Montane Heaths	0.17	0.13	0.00	0.13	0.04	0.04	0.67	0.04	0.04	0.25	0.92	0.29	0.04
Sydney Montane DSF	0.17	0.17	0.00	0.23	0.10	0.07	1.00	0.07	0.00	0.27	0.60	0.60	0.10
South East DSF	0.24	0.30	0.02	0.35	0.37	0.13	0.72	0.09	0.02	0.26	0.50	0.74	0.24
Montane Bogs and Fens	0.23	0.72	0.09	0.11	0.04	0.30	0.13	0.32	0.04	0.06	0.43	0.02	0.04
Montane Lakes	0.09	0.36	0.00	0.05	0.00	0.05	0.00	0.77	0.05	0.00	0.00	0.00	0.00
Central Gorge DSF	0.20	0.14	0.00	0.32	0.18	0.05	0.41	0.09	0.16	0.05	0.18	0.41	0.89

Table 3d: Proportion of species characteristic of each structural/physiognomic class that are diagnostic of units of a 15-cluster solution derived using k-means.). Dark grey indicates proportions>0.7, pale grey proportions >0.5. Cells with the same shading in column one comprise members of the same formation.

Cluster	14	13	15	7	4	5	11	1	2	8	10	6	3
Alpine Herbfields	0.88	0.62	0	0	0.12	0.04	0.04	0.12	0	0.04	0.31	0.04	0.04
Alpine Bogs and Fens	0.43	0.43	0	0.04	0.21	0.11	0.07	0	0	0.04	0.18	0	0.07
Alpine Heaths	0.3	0.89	0	0.11	0.44	0.07	0	0.04	0	0	0.11	0	0
Alpine Fjaeldmarks	0.63	0.88	0	0	0	0	0.06	0.06	0	0	0.06	0	0
Southern Tablelands DSF	0	0	0.74	0.26	0.4	0.31	0.09	0	0.09	0.26	0.03	0	0.57
Southern Escarpment WSF	0	0	0.03	0.97	0.41	0.19	0	0	0.09	0.09	0.03	0	0
Montane Wet Sclerophyll Forests	0	0.09	0.09	0.57	0.8	0.23	0.06	0.03	0.14	0.06	0.14	0	0
Southern Tableland WSF	0	0	0.17	0.56	0.59	0.68	0.12	0	0.12	0.29	0.15	0.05	0.24
Subalpine woodlands	0	0.16	0.27	0.38	0.84	0.49	0.11	0	0.08	0.16	0.14	0.03	0.19
Tableland Clay GW	0.03	0.05	0.08	0.27	0.41	0.62	0.46	0.03	0.03	0.32	0.16	0.08	0.3
Southern Tablelands GW	0	0	0.23	0.23	0.3	0.77	0.53	0	0.05	0.33	0.05	0	0.53
Temperate_Montane_grasslands	0.04	0.04	0.04	0.04	0.07	0.52	0.85	0	0	0.22	0.22	0.19	0.22
Southern Montane Heaths	0	0	0.52	0.07	0.14	0.1	0.03	0.34	0.24	0.07	0.07	0	0.21
Sydney Montane Heaths	0	0	0.13	0.04	0.04	0.04	0	0.88	0.67	0.04	0	0	0.04
Sydney Montane DSF	0	0	0.2	0.1	0.07	0.07	0	0.37	0.8	0.07	0	0	0.1
South East DSF	0	0.02	0.28	0.33	0.15	0.11	0	0.22	0.65	0.24	0.02	0	0.09
Central Gorge DSF	0	0	0.09	0.16	0.02	0.09	0.09	0	0.18	0.86	0	0	0.23
Montane Bogs and Fens	0.4	0.11	0.04	0.04	0.23	0.21	0.06	0.32	0.09	0.02	0.81	0.15	0.04
Montane Lakes	0.23	0.05	0	0	0.05	0.23	0.09	0	0	0	0.73	0.73	0.05

Figure 1: Sample clusters (A-F) Simulated data created by supplying cartesian coordinates for six centroids and generating random coordinates normally distributed around each centroid with sample sizes i) n= 30, 50, 500, 50, 70, 300), and ii) n = 20, 100, 500, 20, 100, 500) with standard deviation =1. The boundaries of each cluster are approximated by circles, colours indicate cluster membership as determined by k-means operating on a matrix of Euclidean distances.

Hosted file

`image1.emf` available at https://authorea.com/users/505959/articles/584876-a-comparison-of-central-tendency-and-interconnectivity-approaches-to-clustering-multivariate-data-with-irregular-structure

Figure 2: Graphical representation of Chameleon's two-phase algorithm (reproduced form Karypis *et al* . 1999)

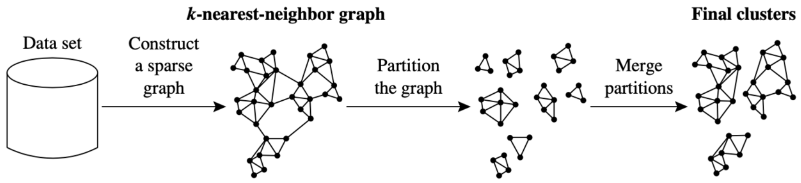

Figure 3: Mis-classification rate (A) and average similarity among objects within clusters (B) function of neighbourhood size as a function of neighbourhood size. Results for solutions obtained with more than

30 sub-partitions are not shown in (A) because samples were frequently concentrated in a single cluster (chaining). Trials incorporating an agglomeration phase (a >15) were performed using a weighted single linkage function.

Hosted file

`image3.emf` available at https://authorea.com/users/505959/articles/584876-a-comparison-of-central-tendency-and-interconnectivity-approaches-to-clustering-multivariate-data-with-irregular-structure

Figure 4: Clustering of simulated data (Figure 1) by Chameleon illustrating increasing rates of mis-classification with increasing neighbourhood size. Samples with the same colour were placed in the same cluster.

Hosted file

`image4.emf` available at https://authorea.com/users/505959/articles/584876-a-comparison-of-central-tendency-and-interconnectivity-approaches-to-clustering-multivariate-data-with-irregular-structure

Figure 5: Trends in mis-classification rate and within-cluster homogeneity with increasing neighbourhood size or increasing number of sub-partitions in the agglomeration phase. The effect of increasing sub-partitions using the single-linkage function is not shown due to chaining as described above). Trendlines are least-squares linear regressions. Data describing the respective 15-cluster solutions derived by k-means, agglomerative and divisive algorithms are plotted for comparison (see figure 6).

Hosted file

`image5.emf` available at https://authorea.com/users/505959/articles/584876-a-comparison-of-central-tendency-and-interconnectivity-approaches-to-clustering-multivariate-data-with-irregular-structure

Figure 6: Trends in mis-classification rate and within-cluster homogeneity with increasing thematic detail (15 – 250 clusters).

Hosted file

`image6.emf` available at https://authorea.com/users/505959/articles/584876-a-comparison-of-central-tendency-and-interconnectivity-approaches-to-clustering-multivariate-data-with-irregular-structure

Figure 7: Cluster sizes ranked in order of increasing size and plotted as a proportion of the number of samples in the largest cluster.

Hosted file

`image7.emf` available at https://authorea.com/users/505959/articles/584876-a-comparison-of-central-tendency-and-interconnectivity-approaches-to-clustering-multivariate-data-with-irregular-structure